DOCUMENT RESUME

ED 400 306 TM 025 665

AUTHOR Brookhart. Susan M.; Masciola, Douglas A.
TITLE The Validity and Reliability of Portfolio Assessment of Eighth Grade Language Arts Students.
PUB DATE 9 Apr 96
NOTE 11p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criteria; Decision Making; Grade 8; Junior High Schools; *Junior High School Students; *Language Arts; Pilot Projects; *Portfolio Assessment; Portfolios (Background Materials); Scoring; Statistical Bias; *Student Placement; Urban Schools; Writing Tests

ABSTRACT

A pilot validity study was conducted of the use of eighth-grade language arts portfolios for ninth-grade English placement decisions (academic or general) in a school district that consists of a city and an independent borough. Portfolios offered an opportunity to collect and examine multiple measures of performance for decisions previously made on the basis of language arts grades. Criteria-for-Placement sheets were designed as cover sheets for the portfolios, and portfolio contents, placement criteria, and rubric design were aligned with the transitional outcomes defined. Writing samples were scored on four dimensions and used with other criteria such as test scores, grades, and work and study habits to place 123 students. The effect of the new criteria was to make recommendations for Academic English more rigorous. The Criteria-for-Placement form exhibited acceptable validity and reliability in the pilot study. Examination of rater bias and of student outcomes after placement are recommended for further study. (Contains eight unnumbered tables and three references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 306

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
SUSAN M. BROOKHART
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

THE VALIDITY AND RELIABILITY OF PORTFOLIO ASSESSMENT OF EIGHTH GRADE LANGUAGE ARTS STUDENTS

Susan M. Brookhart

Douglas A. Masciola

Duquesne University

BEST COPY AVAILABLE

Paper presented at the annual meeting of the

National Council on Measurement in Education

New York, NY

April 9, 1996

TM025665

THE VALIDITY AND RELIABILITY OF PORTFOLIO ASSESSMENT OF EIGHTH GRADE LANGUAGE ARTS STUDENTS

This paper reports the results of a pilot validity study for the use of eighth grade Language Arts portfolio information for ninth grade English placement (Academic or General) decisions in one school district. The district consists of a third class city and one independent borough. Twenty-two percent of the district's student population of approximately 2,100 is African-American, and close to 60% of the students receive free or reduced lunch. The eighth grade language arts classes are representative of the total school population. This is a summative use of portfolios that is teacher-directed and contrasts with both formative and more student-centered uses. Portfolios offered an opportunity to collect and examine multiple measures of student performance for an important placement decision that had previously been made on the basis of Language Arts grades. These results are preliminary in two senses. First, an important piece of information for a validity study is not yet available, namely, the success of students in the Academic and General English classes to which they were assigned. A study examining student success in the 1995-96 school year after placement based on portfolio information is in progress. Second, while the data from this one class of 8th graders look extremely promising, it is not prudent to make definitive conclusions based on one study. A replication study is also in progress; such a replication would make a much stronger case for validity than a single study.

Purposes for Studying the Assessment of Student Performance in the District

Washington School District (WSD) has begun a process of investigating alternative assessments as complements to existing methods of student assessment. The overall purposes for this project are to integrate instruction and assessment, to develop prototype performance assessments that serve as indicators for Pennsylvania academic goals and serve as benchmarks for student progress through the WSD curriculum, and to describe an assessment system in which classroom tests, performance assessments, and state and other secured tests work together to provide evidence of student achievement. During the 1994-95 school year, a pilot performance assessment project focused on the use of eighth grade Language Arts portfolios for placing students into ninth grade English, in either Academic or General classes. The primary purpose for these portfolios was to provide a compilation of evidence for Language Arts achievement that included student writing samples.

Methods

Instrument. Criteria for Placement sheets were designed as cover sheets for portfolios by this process. The two 8th grade Language Arts teachers, the Assistant Superintendent, the Instructional Support Coordinator, and a measurement consultant met for 4 full-day workshop sessions: October 26, 1994, January 18, March 1, and May 1, 1995. During the work days, transitional outcomes for the end of the 8th grade year were matched with WSD curriculum objectives, then to state student learning outcomes. Then the portfolio contents, criteria for placement, and rubric design were aligned with the transitional outcomes. The final draft of the Criteria for Placement sheets were used with the 1994-95 8th grade portfolios to recommend students for placement into 9th grade English classes for 1995-96.

Writing samples were scored on 4 dimensions (Development, Organization, Attention to Audience, and Language), using rubrics on a 1-4 scale (Maryland State Department of Education, 1993). Criteria for Placement sheets listed the following indicators of the 8th grade transitional

outcomes, with placement criteria: (I) descriptive, explanatory, persuasive, and narrative paragraph writing samples, with the criterion of an average score across four dimensions of 3 or 4 on each paragraph; (II) one on-demand composition, with the criterion of 3 or 4; (III) classroom tests, which were selected chapter tests from the English text (D.C. Heath and Company, 1987) chosen for their content match with those transitional outcomes concerning grammar and usage, with the criterion 80% or above on each; (IV) English grade average after the third 9-week report period, with the criterion B or A; (V) Reading grade average after the third 9-week report period, with the criterion B or A; and (VI) work and study habits, with the criterion of an overall 4 average on 4 items rated 1=never through 4=consistently: homework done on a regular basis, demonstrates in-depth reflection on material in class, consistently prepared for participation in class, demonstrates regular attendance.

Teachers completed the sheets and used the decision rule of "meets standard on at least 5 of the 6 criteria" to recommend students for Academic English 9. Students who did not meet at least 5 of the criteria were recommended for General English 9. If the teacher felt a student who did not meet the criteria for recommendation should nevertheless be assigned to Academic English 9, she could so indicate, adding comments at the end of the Criteria for Placement form.

Rater Reliability Data Set. A random sample of 5 persuasive paragraphs per class (15 per teacher, 30 total papers) was drawn. After discussing together the meaning of the rubrics, the two 8th grade teachers independently rated these papers with the 4-point rubric, on the four scales. Interrater reliability and generalizability analyses were done using these data.

Main Data Set. The main data set consisted of all Criteria for Placement Sheets from the 8th grade class of 1995 with complete enough data for teacher placement recommendations, plus standardized achievement measures. Two standardized achievement measures were available: CAT Reading standard scores and Pennsylvania System of School Assessment (PSSA) Reading standard scores. Standardized test scores were concurrent with the portfolio data (PSSA administration was in February, 1995; CAT administration was in April, 1995). Gender, ethnicity, and Language Arts class were also indicated as sorting variables. There were a total of 123 students (41% female, 59% male; 29% African-American, 71% white). Sixty-four students (52%) were placed into General English and 53 (43%) were placed into Academic English; an additional 6 students (5%) were assigned to Academic English.

Results and Discussion

The effect of the Criteria for Placement process was to make recommendation for Academic English more rigorous. The previous year's decision rule for placement into Academic English 9 was a grade of B or better in English. Seventeen students had B's in English after the third 9 weeks but were recommended for General English 9. No students with C or below in English after the third 9 weeks were recommended for Academic English 9, although 2 such students were assigned.

Reliability Questions

1. Are the writing rubrics being used reliably (consistency among raters)?

a. Interrater correlations

Two teachers rated each of 30 papers on the 1-4 rubric for each of four dimensions:

Development (D), Organization (O), Attention to Audience (A), and Language (L). Interrater correlations measure the degree to which each teacher rank-ordered the students in the same way, that is, relative agreement. Correlations between the two teachers' ratings on each of the dimensions were:

INTERRATER CORRELATIONS

r(D) = .77	r(O) = .67	r(A) = .58	r(L) = .71

These correlations were acceptably high except for the dimension of Attention to Audience. These correlations may also be underestimates of the real relationship between the scores because there was not a lot of variability in the ratings.

b. Percent agreement, and percent agreement ±1

The 30 sets of scores were also examined for absolute agreement, that is, the percent of the 30 papers for which the two teachers agreed in assigning a 1, 2, 3, or 4 for each of the dimensions. In addition, the percent of agreement within one point was also calculated.

Dimension	% Agreement	% Agreement ±1
DEVELOPMENT	67	100
ORGANIZATION	63	100
ATTN TO AUD	67	97
LANGUAGE	63	93

While these levels are just acceptable, it would be a good idea to work to raise these levels of agreement. Inspection of the raw data indicated that the disagreements were usually in one direction, that is, where there were disagreements, the same teacher was usually the one to assign the higher score. Mean scores, calculated by averaging the four dimension scores for each student and then averaging scores for each class, indicated that the more lenient rater assigned higher scores to students from her own class, but not to students in the other teacher's class.

MEAN SCORES, BY TEACHER/RATER, WITHIN CLASSES

CLASS ONE

Teacher	N	Mean	Std Dev	Minimum	Maximum
ONE	15	3.7000000	0.4551295	2.5000000	4.0000000
TWO	15	3.2333333	0.6909276	2.0000000	4.0000000

CLASS TWO

Teacher	N	Mean	Std Dev	Minimum	Maximum
ONE	15	2.7500000	0.6477985	1.7500000	3.7500000
TWO	15	2.7666667	0.6010903	1.7500000	3.7500000

This is a common kind of bias and usually results from the teacher who knows the students best seeing the performance of which she knows students are capable, not just immediate performance, in a given work sample. An effective way to combat this bias is usually to simply provide the teachers with information about their scoring and time for them to discuss how they use the rubrics (Herman, Aschacher, & Winters, 1992). If these two teachers receive this report this summer and have a meeting as they plan their portfolio use for the 1995-96 school year, there is a high probability that the scoring discrepancy will not be present in the next year's data.

2. Are students performing consistently?

The statistic *Cronbach's alpha* measures the degree to which students perform in the same relative manner on one item in a composite assessment as they do on the others. To study internal consistency of performance on the Criteria for Placement forms, the six criteria were grouped into four, to eliminate single-item scales. At least two scores are necessary before consistency can be measured. The four areas were: Writing, Tests, Grades, and Work Habits. The Writing scale consisted of the Descriptive, Persuasive, and Narrative writing samples from Criterion I (Explanatory samples were eliminated because one class did not include these pieces in the portfolios), plus the on-demand Composition scored on the same 4-point rubric (Criterion II). The Tests scale consisted of the classroom tests from Criterion III. The Grades scale consisted of English grade (Criterion IV) and Reading grade (Criterion V). The Work Habits scale included the four ratings from Criterion VI: homework, reflection, preparedness, and attendance. Then a stratified alpha was calculated to measure the consistency of student performance on the set of four scales, or the reliability of performance on this group of groups of scores. Alpha levels were acceptable.

SCALE	INTERNAL CONSISTENCY α
WRITING	.64
TESTS	.79
GRADES	.89
WORK HABITS	.95
TOTAL	.79

The lower alpha for Writing reflects the fact that the on-demand composition scores were different from the classroom writing samples. These two kinds of writing appeared to be tapping somewhat different groups of skills. The correlation between Composition score and total Writing score was .28 (compared with .61 for the Descriptive paragraphs), and the estimated reliability for the Writing scale without the Composition score was .69, higher than the .64 value obtained for all the Writing samples together. On the Criteria for Placement form, the Composition score is considered separately, and was only joined with the other Writing samples in this study for the purpose of calculating a total reliability for the portfolios. The overall reliability of .79 was acceptable.

3. What is the generalizability of student performance across raters and task domain representation, when these are considered simultaneously?

The small data set of 30 papers that were graded on four dimensions of writing by two raters each allowed the possibility of a generalizability study with the design Persons X Raters X Dimensions. A generalizability study examines the contribution to total variance in performance

made by each of the facets. Reliability (generalizability) is high when the variance due to persons (that is, variations in students' performances) is high relative to sources of variance that represent error in measurement (like rater variation) or sources of variance across items (in this case dimensions). Decision studies based on estimated variance components allow for estimating the reliability (generalizability) that one could expect with different numbers of raters or tasks. Of particular interest here is estimated reliability for one rater, which is the usual case for judging student writing samples.

GENERALIZABILITY STUDY, PERSONS X RATERS X DIMENSIONS

SOURCE OF VARIATION	EST. VARIANCE COMPONENT	% TOTAL VARIATION
PERSONS	.3995	59.6
RATERS	.0207	3.1
DIMENSIONS	.0443	6.6
PERSONS X RATERS	.0654	9.8
PERSONS X DIMENSIONS	.0140	2.1
RATERS X DIMENSIONS	.0057	0.1
PERS X RATERS X DIM	.1206	18.0

Notice that the variance due to persons is the largest share of variation. This is the reason for the acceptably high levels of generalizability reported in the following table. The Persons X Raters interaction component will be reduced when the rater bias discussed in section 1 is reduced.

Estimated reliability coefficients for various combinations of raters and numbers of dimensional scores are presented in the table below. Relative decisions are ranking and grouping decisions, where students' performance relative to one another is what is being measured. Absolute decisions are decisions about absolute levels of performance. Both of these decisions are relevant to the present study. The placement decisions about 9th grade English can be considered relative decisions, because they allocate available space, although one could argue for considering this an absolute decision. Reports about progress on state outcomes can be considered absolute decisions, in that they are meant to report what the students can do.

DECISION STUDY, ESTIMATED GENERALIZABILITY FOR 1-2 RATERS AND 1-4 DIMENSIONS, CONSIDERING DIMENSIONS AS A FIXED FACET

# RATERS	# DIMENS	GENERALIZABILITY FOR RELATIVE DECISIONS	GENERALIZABILITY FOR ABSOLUTE DECISIONS
1	1	.67	.61
1	2	.76	.71
1	3	.79	.75
1	4	.81	.77
2	1	.80	.73
2	2	.86	.82
2	3	.88	.85
2	4	.89	.87

Two conclusions can be drawn form the Decision Study table. First, the level of generalizability is acceptable using any number of dimensions if there are 2 raters, and acceptable

for two or more dimensions using one rater. Second, adding a rater improves reliability more than adding a dimension.

Validity Questions

1. Does the Criteria for Placement Form have content validity?

a. How does the assessment fit with the curriculum?
b. How does the assessment fit with the state outcomes?

During 4 day-long work sessions, the team of teachers, administrators, and consultant finalized the content of the Criteria for Placement forms. The forms were adapted from an earlier version by comparing them with the 8th and 9th grade curriculum objectives, with transitional outcomes (middle school to high school) prepared by the district, and by discussing the skills needed for success in the instructional activities that comprise instructional delivery in 9th grade English. Items were deleted from the list if they did not match the curriculum, and items were added to address deficiencies in the list, so that the result conformed to both the district and state outcomes. The reader is referred to lists of these outcomes in the District Office for specific details. Workshop participants also discussed the standards of performance that should be expected for each criterion. The most important consideration was the level of skill required to succeed in the high school classes, at the level of a B average. After several drafts, the Criteria for Placement form (see appendix) was judged ready for pilot testing. It may be revised in the future as curriculum changes are made.

2. Does the Criteria for Placement Form have construct validity?

The evidence for content validity derived from the development process, described in section 1 above, is relevant to construct validity. The "constructs" that the Criteria for Placement form seeks to measure are classroom-related achievement constructs, specifically related to the writing process and the grammar and usage concepts that are building blocks for writing. Thus alignment with the outcomes intended for Language Arts classroom instruction is important.

To further examine the relationships among the measures on the Criteria for Placement Form, a factor analysis was done. CAT Reading score, PSSA Reading score, and summary scores on Criteria I through VI were factor-analyzed to identify latent variables or underlying, unmeasured constructs. The purpose for this analysis was to see whether the classroom-related achievement measures presented in the portfolios and one the Criteria for Placement forms were providing information beyond that available in the standardized achievement measures. If the multiple measures were providing information redundant to that from the tests, then test scores would be a more efficient way to get the information. It was expected, however, that the factor analysis would demonstrate that classroom achievement represented a different underlying construct.

Two factors accounted for all of the common variance in the 8 scores (two tests and six criteria). These factors could be interpreted as a classroom work/achievement factor (Factor 1) and a verbal ability factor (Factor 2). These two factors were highly related ($r=.82$) but distinct. Classroom writing, classroom tests, grades, and work habits all loaded on Factor 1. Grades, classroom tests, and work habits most clearly defined this construct. The standardized tests, classroom writing, and the on-demand composition loaded on Factor 2. Standardized tests most

clearly defined this construct. The factor analyses gave evidence that the classroom-based measures included as Criteria for Placement measured underlying factors that were different from the standardized testing measures.

FACTOR ANALYSIS RESULTS*

Rotated Factor Pattern (Std Reg Coefs)

	FACTOR1	FACTOR2
CAT	-0.15	0.99
PSSA	-0.01	0.91
I-WRITING SAMPLES	0.47	0.34
II-ON-DEMAND COMP	0.28	0.33
III-CLASSROOM TESTS	0.88	-0.03
IV-ENGLISH GRADE	1.12	-0.24
V-READING GRADE	0.69	0.20
VI-WORK HABITS	0.82	0.07

* Principal axis factor analysis, priors=squared multiple correlations, Harris-Kaiser rotation

The loadings also suggest that the on-demand composition was as representative of the underlying construct measured by standardized tests as it was of classroom achievement. The classroom writing samples represented both of these constructs but was more closely associated with classroom achievement.

3. Is the Criteria for Placement Form free from bias?

Two bias studies were performed, one for gender and one for race. The questions investigated were:

a. Do girls and boys with equivalent abilities have an equal likelihood of being placed (recommended or assigned) in Academic English 9?

a. Do white and black students with equivalent abilities have an equal likelihood of being placed (recommended or assigned) in Academic English 9?

CAT Reading score was used as a proxy measure for "ability." Contingency table analyses checked to see whether there was a significant difference in proportion of females/males or white/black students, respectively, placed in academic English 9 at each CAT score level.

CONTINGENCY TABLE ANALYSIS OF PLACEMENT BY GENDER AND RACE, CONTROLLING FOR READING ABILITY

FACTOR	MANTEL-HAENSZEL χ^2	p
GENDER	.33	.57
RACE	.36	.55

The lack of significant difference for gender or race suggest the answer "yes" to questions A and B above. Girls and boys of equal ability have chances of being placed in Academic English 9 that do not differ significantly (that is, beyond what would be expected by chance). White and black students have similarly equal chances, given ability. The portfolio placement process is free from bias defined in this statistical sense.

The contingency table analysis also permitted calculating odds ratios for gender and race. Odds ratios compared the odds that girls and boys, or white and black students, of equal abilities have of being placed in Academic English. The odds ratio for Gender was 1.692; a boy at a given ability level had odds of being placed that were 1.7 times the odds for a girl. The odds ratio for Race was 1.5; a white student at a given ability level had odds of being placed that were 1.5 times the odds for a black student. These odds ratios were, as demonstrated above, within the chance range. Estimates for Race are tentative because there were not enough students of both races at all ability levels to form complete contingency tables.

Conclusions

The Criteria for Placement form exhibited acceptable validity and reliability in the 1995 pilot study. Because of the preliminary nature of the findings, this conclusion should remain tentative, awaiting the results of further study. The 1995-96 study should replicate this pilot study and extend it. Specifically, additions for 1995-96 include two questions.

Can the rater bias be removed? After a meeting prior to the next set of scoring, the 8th grade teachers should move closer to congruence in their ratings. Even in the pilot study, the two teachers did an excellent job, resulting in acceptable levels of reliability calculated in three different ways. Nevertheless, an effort should be made to even out the treatment of the two classes by the different raters. Reliability coefficients would then rise even higher. More importantly, student writing scores would be less dependent on class. This problem can be resolved for aggregated scores by using weights, so that for example a corrected 8th-grade average writing score could be reported to the state as an indicator of a transitional outcome. But the bias remains a problem if the individual portfolios are treated with a decision rule to determine placement in 9th grade English. Weighted corrected scores are not feasible at the individual level. It would be most helpful, too, to add an external rater in the next rater reliability study, both to add a measure of objectivity and to see how easily papers can be rated by other teachers, community members, etc., who might some day become partners in the process. The two teachers in the pilot study were unusual in their close coordination of classroom practices. Before becoming institutionalized, the Criteria for Placement forms and their portfolios should prove themselves workable for most teachers.

How do the students fare in their 1995-96 placements? Do the students succeed in their Academic and General English sections, respectively? Both their achievement levels and satisfactions will be investigated. Teacher judgment about the success of the placement process will also be investigated. Special attention will be paid to those whose placements differed from what they would have been under the previous decision rule: the 17 B students in 8th grade English who were placed in English 9. Their successes in particular would be compelling evidence for the validity of the portfolio placement criteria.

There are other validity questions that could be addressed as the district continues the placement process year after year. Probably the most important one has to do with instructional

validity. What classroom instruction decisions do teachers make on the basis of the portfolios? Are they satisfied with the quality of the information? How does work on the 8th grade portfolios prepare students for success on 9th grade writing, classroom projects, tests, etc.?

The factor analysis gave preliminary evidence that classroom success is related to, but distinct from, student performance on secured tests. In today's climate of accountability, it would be interesting to investigate the nature of this relationship further. Particularly, it would benefit school and community if demonstrations of student achievement beyond test performance were described and exhibited. Bringing in outside readers for student writing might begin this process.

The many questions in this conclusion section are offered in the spirit of recommendations for next steps. This report should close by returning to the main conclusion, namely, that the pilot study found acceptable validity and reliability for all measures, even when there were some problems identified. This was a surprise to the authors, because much performance assessment data in the literature has been reported to behave in a messy and inconsistent fashion--typically one reads that the "technical quality" of the performance assessments is suspect. This was not the case in the Washington School District data that formed the basis for this report.

References

Heath Grammar and Composition 8. (1987). Lexington, MA: D. C. Heath and Company.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Maryland State Department of Education. (1993). *Scoring rubrics.* Unpublished papers.